Appunti di Archivismi 639/ Rinfrescare, riordinare, rivedere, rimpolpare

(639)—Quando si crea un'archiviazione di informazioni se ne deve curare anche aggiornamento, manutenzione e correzioni. E se si è usato…

Appunti di Archivismi 639/ Rinfrescare, riordinare, rivedere, rimpolpare

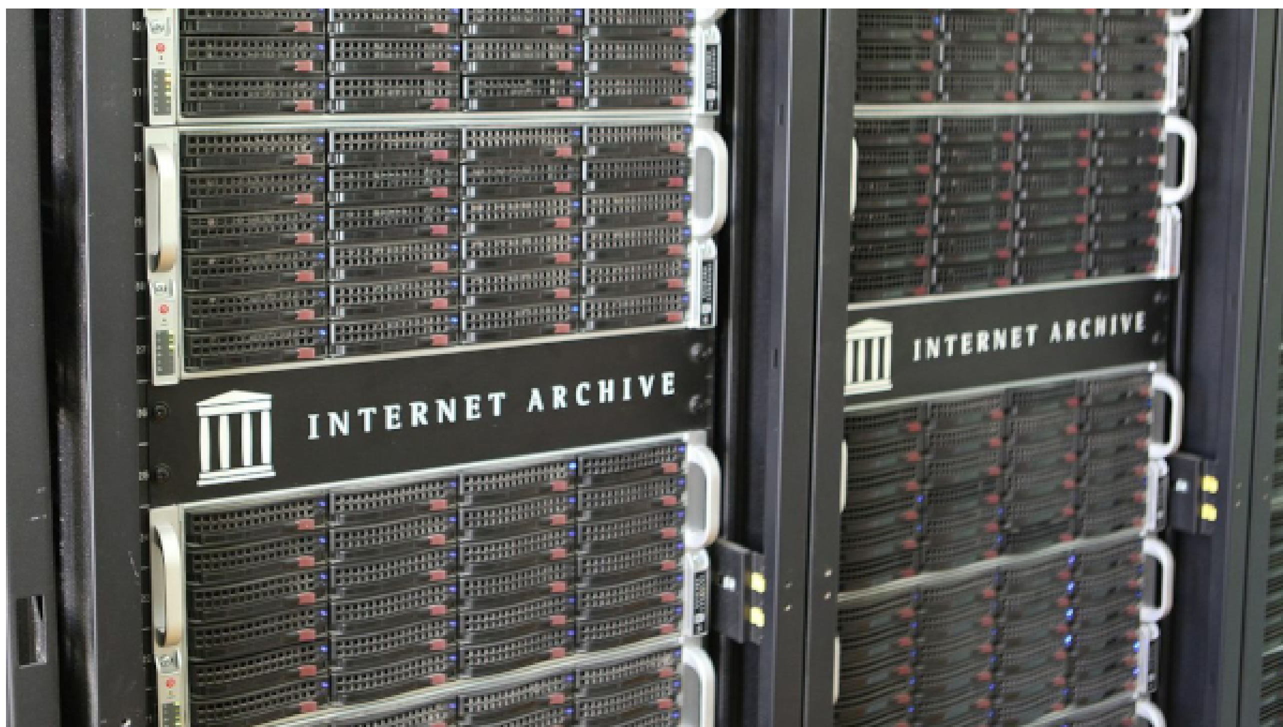


Figure 1:

(639)—Quando si crea un'archiviazione di informazioni se ne deve curare anche aggiornamento, manutenzione e correzioni. E se si è usato Internet Archive è bene conoscere potenzialità, problemi e trucchi

26 Agosto 2025—L'iniziativa "*Archivismi*" ha lasciato a Cassandra l'onere di occuparsi (con gioia) di tutto quanto archiviato.

E se l'archiviazione nell'*Arctic World Archive* non permette, e quindi neppure richiede, nessuna manutenzione, essendo laggiù possibile solo effettuare una nuova archiviazione, altrettanto non si può dire per le archiviazioni fatte su Internet Archive.

Si tratta infatti di archiviazioni "vive", come in una biblioteca vera, si possono inserire nuovi libri, si può "cancellare" un libro danneggiato, si possono aggiornare i libri con un'edizione successiva qualitativamente migliore, si possono aggiungere allegati ad un libro, si possono ampliare o correggere collezioni di libri.

Come in precedenti articoli di Archivismi, anche questo sarà formato da singoli punti monotematici; se quello che vi è descritto vi dovesse capitare, avrete così una soluzione già pronta in mano. E comunque anche solo leggerli serve anche a capire come muoversi.

Zero: Grazie al lavoro fatto, **Cassandra ha scoperto una grossa falla nel processo di estrazione dei dati che Medium.com offre**, cosa che solo per fortuna non ha prodotto serie conseguenze nell'archiviazione di Cassandra Crossing.

Quando si esportano i dati di un account di Medium.com, si ottiene, come in altri social, una serie di directory, ciascuna delle quali contiene informazioni di un certo tipo, il tutto eventualmente con una pagina html di indice.

Questo era il caso di Medium.com; scompattando l'estrazione, una della directory conteneva tutti gli articoli in formato html. Si trattava però di un html pessimo, pieno di tag utili solo al funzionamento del loro sito, che lo rendevano illeggibile e difficilmente elaborabile.

Quindi, prima di produrre i file per l'archiviazione, era necessario "ripulire" l'html, e solo poi generare i file markdown, html e pdf destinati all'archiviazione finale.

Tutto questo era stato eseguito in maniera automatica da un singolo script bash, tramite l'impiego di normali utility unix e di Pandoc, un programma di conversione di formati testuali estremamente potente, che usa LaTeX come formato intermedio. Modificare lo script per altri sistemi operativi dovrebbe essere fattibile, ma lasciamo volentieri l'esercizio al lettore, che nel frattempo potrebbe anche cadere vittima della tentazione di passare ad un sistema operativo libero.

Cassandra vi aveva già fornito e spiegato lo script nelle precedenti puntate di Archivismi, in particolare questa e questa.

L'ultima versione, come quella di allora, non è perfetta ma sempre un work-in-progress; Cassandra ve lo riallega nuovamente, per i piccoli ma importantissimi cambiamenti che ha apportato in questi mesi (e questa volta non c'è bisogno di convertire i caratteri!).

```
# Procedura per la preparazione all'archiviazione degli articoli
# di Cassandra Crossing
#
# inizializzazioni varie
_base="./"
_base2="./posts/"
_base3="./markdown/"
_base4="./temp/"
_base5="./html/"
_base6="./pdf/"
_base7="./post2/"
_base8="./md2/"
_temp="temp.txt"
_temp2="temp2.txt"
_temp3="temp3.txt"
#
# cambio directory di lavoro, creazione directory e pulizia file
cd "${_base}"
mkdir markdown html temp pdf post2 md2
rm ./markdown/* ./html/* ./temp/* ./pdf/* ./post/temp*  ./post2/* ./md2/*
cd "${_base2}"
rm "${_temp}"  "${_temp2}"
_dfiles="*"
#
# inizio loop principale
```

```
for f in $_dfiles
do
rm "${_temp}"
#
# estrazione del numero dell'articolo
g=`grep -Eo -m 1 '\([0-9]+\)' $f | tr -d '()'`
g="000"$g
g=`echo $g | rev | cut -c 1-3 | rev`
h=`echo $f | cut -d '_' -f2- | rev | cut -d '-' -f2-| rev`
#
# formazione del nuovo nome del file e copia col nuovo nome
i=$g"_"$h
echo "---> Identifier: $i"
echo "$i" >> "${_temp3}"
cp $f "../$_base4${i}.html"
cp $f "../$_base7${i}.html"
#
# conversione in formato markdown, ripulitura e riconversione in html

# traduzione del file html da Medium.com in markdown con risorse incorporate
pandoc --embed-resources=true -f html -t markdown  "../"$_base4$i".html" > "${_temp}"

# pulizia del markdown da porcherie html provenienti da Medium.com
grep -v "^:::" "${_temp}" |sed -e 's|$i".md"

# traduzione del markdown embedded in html embedded
pandoc -V geometry:margin=2cm --embed-resources --standalone -f markdown -t html   "../'

# traduzione html in pdf, con size, pagina, link e margini, ma sposta le immagini grandi
pandoc --metadata title="" -V papersize:a4 -V fontsize=12pt -V colorlinks   -V geometry:

echo `cat "../"${_base3}$i".md" | tr "\n" " " | sed -e "s/.*\*(//" | sed -e "s/\*.*//"`
#
# pulizia e fine ciclo
done
#rm -rf "${_temp}" "../$_base4"
```

All'inizio Cassandra aveva deciso di archiviare il file html originali provenienti da Medium.com come fonte "originale" degli articoli, considerando gli altri file prodotti da questi come file "derivati". Si era accorta però che il file originali non contenevano le immagini, ma solo dei link che puntavano a server di Medium.com, cosa abbastanza comune.

Aveva quindi fatto in modo di archiviare anche una seconda versione di HTML che, grazie ai miracoli di Pandoc, conteneva le immagini come risorse codificate rot64.

Durante l'ultima campagna di aggiornamento della collezione degli articoli di Cassandra Crossing, la vostra profetessa preferita si è tuttavia accorta **che molti degli articoli originali erano troppo piccoli per la quantità di testo che avrebbero dovuto contenere.**

Ha constatato, non senza una buona dose di raccapriccio e di rabbia verso chi "bara" nei confronti dei propri utenti, che **intere parti di testo, e spesso l'intero articolo, non erano contenute nel file ma venivano anche esse caricate come risorse esterne dai**

server di Medium.com.

Insomma, nei file “originali” non solo le immagini ma anche intere parti di testo degli articoli non c’erano, ma erano, per fortuna, state incluse successivamente dal processo di conversione realizzato da Cassandra.

Andreottianamente, Cassandra non crede che queste “caratteristiche” siano state inserite per sbaglio.

Morale: Cassandra ha deciso che il file “originale” degli articoli sarà un file markdown con le immagini embedded, e che il file html di Medium.com continuerà ad essere allegato solo per motivi “storici” come file accessorio.

Questo richiederà una riarchiviazione completa di tutti i file principali di tutti articoli, anche se non richiederà la creazione di nuovi oggetti e nemmeno la modifica dei metadati. Gli automatismi di Internet Archive e lo script di archiviazione dovrebbero permettere di risolvere il problema con un paio di ore di lavoro.

Uno: il formato dei pdf per l’archiviazione e la leggibilità dei suoi contenuti sono probabilmente la cosa più importante per l’utente.

La prima archiviazione di Cassandra Crossing su Internet Archive era stata fatta dopo aver superato molti ostacoli, e con una conoscenza molto limitata di Pandoc e LaTeX.

Per questo i pdf prodotti erano ben lungi da essere ottimali, perché l’uso di Pandoc non permetteva di cambiarne le caratteristiche in maniera semplice, come si sarebbe fatto con qualunque elaboratore di testi. I primi pdf avevano i seguenti problemi:

- [margini ridicolmente larghi;]
- [font eccessivamente piccoli;]
- [la prima pagina di ogni articolo aveva una bruttissima intestazione, contenente l’ID Internet Archive del documento.]
- [gli articoli più recenti iniziavano con una doppia copia di titolo e sottotitolo;]
- [i link erano inclusi nel pdf e cliccabili, ma la loro veste grafica era assolutamente indistinguibile dallo scritto normale, quindi per trovarli ed utilizzarli, bisognava passarci sopra il cursore per vedere quando cambiava da freccia a manina;]

Risolvere tutti questi apparentemente banali problemi ed ottimizzare alcuni altri aspetti, ha richiesto un minimo di studio di Pandoc e molte, ma davvero molte prove; le modifiche si sono alla fine tradotte nell'aggiunta di pochi parametri sulla linea comandi di pandoc.

Tuttavia la risoluzione di questi problemi ha comportato dover lasciare maggiormente la “mano libera” a LaTeX, che ha preso l’autonoma ed inarrestabile decisione di spostare le immagini troppo grandi in posizione fuori testo.

Cassandra leva perciò il suo grido di aiuto a chi potesse aiutare a risolvere il problema, partendo per esempio dal confronto dei file html e pdf di un articolo. Grazie anticipatamente!!!

Cosa dire per tirare un po’ le somme di questa esternazione? Probabilmente la cosa più importante, almeno per chi deve archiviare molti dati, è la **raccomandazione di stare estremamente attenti ai dati che vi vengono forniti dalle applicazioni cloud, al loro formato, alla loro completezza ed alla presenza di risorse remote**.

Anche Cassandra, notoriamente pignola, ci stava cascando.

Due: la gestione delle correzioni necessarie ad alcuni degli oggetti preesistenti, creati in maniera errata (pochissimi, per fortuna), ha richiesto abbastanza tempo e non è ancora stata completata;

per questa ragione Cassandra la rimanda al prossimo numero di questa piccola saga.

Resta comunque a disposizione dei suoi lettori per consiglli o discussione al suo abituale indirizzo di email cassandra@cassandracrossing.org.

Enjoy!

Scrivere a Cassandra—Twitter—Mastodon
Videorubrica "Quattro chiacchiere con Cassandra"
Lo Slog (Static Blog) di Cassandra
L'archivio di Cassandra: scuola, formazione e pensiero

***Licenza d'utilizzo**: i contenuti di questo articolo, dove non diversamente indicato, sono sotto licenza Creative Commons Attribuzione—Condividi allo stesso modo 4.0 Internazionale (CC BY-SA 4.0),* tutte le informazioni di utilizzo del materiale sono disponibili a questo link*.*

By Marco A. L. Calamari on August 27, 2025.

Canonical link

Exported from Medium on August 27, 2025.